Investigating Unfolded Proteins by Small-Angle X-Ray Scattering

by

Danna Li

Department of Chemistry
Duke University

Date:_____

Approved:

_____
Terrence G.Oas, Supervisor

_____
Qiu Wang

_____
David Richardson

_____
Patrick Charbonneau

Thesis submitted in partial fulfillment of
the requirements for the degree of Master of Science in the Department of Chemistry in the Graduate School
of Duke University

2013

ABSTRACT

Investigating Unfolded Proteins by Small-Angle X-Ray Scattering

by

Danna Li

Department of Chemistry
Duke University

Date:_____

Approved:

_____
Terrence G.Oas, Supervisor

_____
Qiu Wang

_____
David Richardson

_____
Patrick Charbonneau

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Chemistry in the Graduate School of Duke University

2013

Abstract

A clear description of the unfolded state is important for understanding protein folding/misfolding reactions. In addition to general ensemble-averaged properties, distributional residue-specific information is particularly necessary for identifying the molecular causes of many protein misfolding diseases. To this end, an anomalous SAXS (small angle X-ray scattering) technique was developed that provides residue-to-residue distance distribution information for unfolded proteins under physiological conditions. A peptide corresponding in sequence to the first helix of λ repressor was used for preliminary experiments with the proposed technique. Selenium and mercury labels were attached to the termini of the peptide and SAXS data of the labeled peptide were collected at the Argonne National Laboratory. End-to-end distance distribution for selenium-labeled peptide was obtained and the viability of the method was discussed based on experimental and simulation results.

Contents

List of Figures

Acknowledgements

I am not dedicating this book to anyone, because its contents are not good enough to be presented to the ones I love or the ones who love me.

However, I do feel indebted to numerous people including my family members, friends, teachers and colleagues.

I owe special thanks to my parents, who always granted me considerable freedom to make important decisions and to choose my own path. I would also like to thank my grandfather Qingyun Shi, who shared with me many interesting views regarding science, art and everyday life. I wouldn't have chosen chemistry without you.

I am grateful to my friends here at Duke: It is you who made my graduate school life more enjoyable. In particular, I would like to thank Zhao Zha, who has been and will continue to be a perfect companion. I wish I could dedicate this book to you when I was halfway through my project; only it does not turn out to be as fascinating as your personality. I feel lucky to have met you.

I thank Dr. Oas for being a wonderful advisor. I still remember the interesting discussions we had during the meetings. You shared with us countless ideas about science, academic writing and many other aspects. I really enjoyed hearing and thinking about them.

I also thank Dr.Wang, Dr.Richardson and Dr. Charbonneau. I benefitted a lot from your classes as well as your suggestions during my preliminary exam.

Finally, I would like to thank my colleagues at Oas lab including Jonathan, Billy, Roy, Jo Anna, Yang, Shiwen, Peifen, Kyle, Pamela and Andrew, who have all helped me to made my life easier in the lab. In particular, Jonathan, Billy and Jo Anna taught me about microbiology which I knew little about before joining the group. Jo Anna guided me through SAXS theory and experiments, and we had several enlightening discussions via email. Roy always explains in detail in response to my theoretical questions, and he offered nice suggestions of my prelim documents and presentation.

I am truly grateful to all of you.

Introduction

Proper folding of most proteins is crucial to their biological function. Protein misfolding, on the other hand, may become pathogenic and is in fact a process underlying many neurodegenerative diseases. Parkinson's disease, for example, is caused by the misfolding of an intrinsically unfolded protein called α-synuclein (αSN) which forms cytotoxic aggregates that leads to massive death of nerve cells [1]. As the precursor of the toxic αSN aggregates, the unfolded αSN acts as the starting point of the misfolding process; from a chemist's view, it is the reactant of the misfolding reaction. Therefore, a clear understanding of the unfolded state will provide rich information about the origin of the misfolding process.

To date, many studies have been directed at understanding the unfolded state of proteins, most of which only succeeded in giving the ensemble-average properties of the unfolded protein. However, the ensemble of conformations that make up an unfolded protein is very broad and therefore average structural information is particularly uninformative. In order to describe the exact causes of the misfolding, it is essential for us to develop a model that captures the essentials of the unfolded ensemble. On the experimental side, we need a method that can determine distributional structural information in a site-specific way. Furthermore, since unfolded proteins stay in solution

under physiological conditions, this method should be amenable to solution-phase measurements.

To this end, I developed an anomalous SAXS (small-angle X-ray scattering) technique ("anomalous SAXS" refers to a particular X-ray scattering phenomenon explained later) that allows us to determine the structure of an unfolded peptide in solution. In particular, information of the residue-to-residue distance distribution is shown to be potentially obtainable by labeling the peptide with selenium.

In Chapter I, a coarse-grained unfolded state model (the helix-coil polymer physics model, abbreviated as HCPP) is described which applies statistical mechanics to the unfolded ensemble. This model serves as theoretical background for understanding the meaning of our experiments and a guide for designing new experiments.

In Chapter II, the model peptide LRH1x (Lambda Repressor Helix I) used for developing the ASAXS technique is discussed. It corresponds in sequence to the first helix of Lambda Repressor, and it possesses many properties desirable for my study. This chapter also introduces the selenium and mercury labeled LRH1x, as they will be used for important SAXS studies in the following chapter.

Basic principles of SAXS and ASAXS are discussed in Chapter III. This serves as the theoretical foundation for understanding the experimental technique that I developed. Following that, the application of selenium label and mercury label in this technique are described. Each of the two labels has its own strengths and weaknesses.

Through an indirect Fourier transform, the residue-to-residue distance distribution of LRH1x was obtained from the ASAXS data. However, concerns regarding the low SNR in these experiments are expressed, which cast doubt upon the reliability of the distance distribution.

Finally, in Chapter IV, I reflect upon the ASAXS technique and propose future directions. I simulated ASAXS curves for doubly-labeled using distance distributions resulted from the combination of many different Gaussian distributions. The results show that the ASAXS curve is quite sensitive to the shape of the distance distribution, meaning that our method is promising in providing useful information of the unfolded ensemble.

I hope that this book will enhance the readers' understanding of the capabilities and limitations of SAXS through the successes and failures of the experiments discussed here. Perhaps the most prominent feature of the unfolded protein is its complexity, which poses a tough test for experimental designers. In the Conclusion section, meanings and implications of this feature are discussed. The author is quite optimistic about future research in this area regardless of the let-downs of her own results.

1. Helix-coil and Polymer Physics theory

1.1 Traditional helix-coil (HC) theory

The "structure" of the unfolded protein can best be characterized by a statistical ensemble (a collection of conformers). Since proteins consist of up to thousands of residues and the (ϕ, ψ) angles of each residue may take a broad range of possible values, the total number of conformations available to the unfolded protein is usually supra-astronomical, making these conformations impossible to enumerate completely.

Helix-coil theory [2], on the other hand, provides a simplification to the enumeration problem while retaining the essential features of the unfolded state. It evaluates protein conformation at the residue level as opposed to the atomistic level. As a further simplification, each residue can be in either of the two states: the helix state (h) and the coil state (c). A residue is in the helix state if its (ϕ, ψ) values fall into the helix region on the Ramachandran plot, and is in the coil state if it is not in the helix state. This binary division of the Ramachandran plot can be justified by the considerable free energy barrier between these two states, which leads to low populations of the intermediate states. By construction, 2^k conformations are available to a k-mer (a protein consisting of k residues), giving us a much more tractable number to enumerate, although most implementations of helix-coil theory do not do this.

In the unfolded ensemble, each conformation is represented by a list of h's and c's. For example, (hhhhh) and (hhhhc) are two possible conformations for a 5-mer. Since

h and c have distinct Gibbs free energies, different conformations consequently have different free energies, leading to their different populations within the ensemble. From an exhaustive enumeration of the ensemble members and the determination of their relative populations, various ensemble-averaged properties of the ensemble can be calculated.

Previously, the helix-coil theory has primarily been used to predict ensemble- and chain-averaged properties such as CD spectra of partially helical peptides [2]. However, average properties are incapable of providing residue- or segment-specific information that are necessary for understanding the fine but important features of the unfolded ensemble. The behavior of specific segments or residues within the protein is still elusive without distributional information such as residue-to-residue distance distributions, which can be obtained from anomalous SAXS data. These details are necessary in order for us to identify the origin of protein misfolding, and are crucial for devising treatment on the molecular level.

In order to investigate the detailed properties, a description of the lengths of helix and coil segments of the unfolded state conformers is needed. While the length of helical segments are fairly easy to compute because the helical geometry is known, the coil part is not yet clearly defined in the traditional helix-coil theory, leaving a gap between the existing helix-coil model and the prediction of residue-to-residue distance distributions.

1.2 Mathematical description of HC theory

In the helix-coil model, a residue in a given conformer can be in either the helix state (h) or the coil state (c). There are two types of helix states: the helix states with one or two coil neighbors (Type I), and those with two helical neighbors (Type II). Setting coil as the reference state with a statistical weight of one, we can then write the statistical weight of the two helix states:

$$v = \exp[-\beta(-T\Delta S_R] \quad \text{(Type I)}$$

$$w = \exp[-\beta(\Delta H - T\Delta S_R] \quad \text{(Type II)}$$

v and w correspond to the probability of these two states relative to the coil state, respectively. Here, v is the helix nucleation parameter, which describes the probability of initiating the formation of a helical segment. v is usually smaller than one, since there is a considerable entropy price to pay in this process. w is the helix extension parameter, it is usually larger than one due to the cooperativity in helix formation. Cooperativity arises because the enthalpic contribution made by several successive helical residues is needed in order to compensate for the initial entropy loss.

ΔH is the negative enthalpic change in helical formation that is considered to be sidechain independent since it's mainly contributed by backbone hydrogen bond formation. ΔS_R is the entropy loss in this process; it is sidechain dependent because it includes both the backbone and sidechain entropy change. ΔH is not included in the

expression of v, because no H-bond is formed when a residue becomes helical and is flanked by two coil residues.

Take a 10-mer as an example. One possible conformation for such a peptide is (cchhcchhhc), the state of which can be represented as (uuvvuuvwvu) according to the rules explained above. The probability of this state, or the population of this conformer, is then given by u^5v^4w/Z, where Z is the partition function. This partition function can be calculated via matrix multiplication, where the matrix is written as

$$\mathbf{M}_j = \begin{array}{c} \\ h\underline{h} \\ h\underline{c} \\ c\underline{h} \\ c\underline{c} \end{array} \begin{array}{c} \begin{array}{cccc} \underline{h}h & \underline{h}c & \underline{c}h & \underline{c}c \end{array} \\ \begin{bmatrix} w_j & v_j & 0 & 0 \\ 0 & 0 & 1 & 1 \\ v_j & v_j & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array}$$

where w_j and v_j are the w and v parameters for the j^{th} residue, and the row and column indices indicate the state of residue (h or c) to its left and right, respectively. The letters underlined represent the state of the j^{th} residue. Take element [3, 2]** as an example: the j^{th} residue is in the helix state as indicated by the underlined "h", and its neighboring residues are both in the coil state as indicated by "c<u>h</u>" and "<u>h</u>c". Therefore, the j^{th} residue is in Type I helix state, whose statistical weight is v, represented by the "v_j" entry in the matrix. The zero entries fill positions that cannot exist in practice. For instance, element [2, 1] is zero because the j^{th} residue cannot be in the coil state (indicated by "h<u>c</u>") and helix state (indicated by "<u>h</u>h") at the same time. The row vector and the column vector at the termini are for the two end residues; the 0's and 1's are arranged in this particular

way to reflect the fact that there is no helical residue to the left and right of the peptide chain.

The matrix multiplication can be written as:

$$Z = (0, 0, 1, 1)\left(\prod_{j=2}^{l-1} \mathbf{M}_j\right)(0, 1, 0, 1)^T$$

This expression generates all combinations of h's and c's with the correlation of states preserved in the product by proper multiplication of rows and columns, which gives the partition function Z. Then the relative population of each conformer can be determined [2].

By taking into account factors like the capping effect and sidechain interactions, the $\mathbf{M}_j$ matrix has been modified and expanded in various ways. Experimental conditions, like pH and temperature, can also be accounted for in this model. For example, the temperature dependence of ΔS and ΔH can be calculated using the AGADIR algorithm [1] in which

$$\Delta H(T) = \Delta H_0 + \Delta C_P\,(T - T_0)$$

$$\Delta S(T) = \Delta S_0 + \Delta C_P \log(T/T_0)$$

where the reference temperature T_0 is 273 K at which ΔH_0 and ΔS_0 are measured. ΔC_P is the change in heat capacity assumed to be sidechain-independent.

Using the HC model, we can calculate the ensemble composition under a given set of experimental conditions. The ensemble composition here refers to the distribution

of all possible helix-coil conformers. For example, the ensemble composition for a two residue peptide could be {45% cc, 30% hc, 20% ch, 5% hh}. Changing experimental conditions would shift this distribution accordingly, which can also be predicted by HC theory.

1.3 Introducing polymer physics in HC theory

Polymer physics can be regarded as a branch of applied statistics, where the tools of statistics are used to study a complex system of polymers. Although it has been used in several studies on chemically-synthesized as well as biological macromolecules, polymer physics has rarely been merged with the helix-coil theory for investigating the coil segments in the unfolded ensemble.

Compared with atomic simulations, models from polymer physics "can capture the essence of the distribution functions without being bogged down by atomic details" [3]. Therefore, joining polymer theory with traditional helix-coil theory will presumably produce a desirable coarse-grained model for the unfolded state ensemble.

There exist many kinds of polymer models. Three basic parameters are generally involved in these models: bond length, bond angle and torsion angle (Fig.1). Each polymer model makes different assumptions about the allowed values of these three parameters [4]. The simplest case is a random-flight chain, where the bond length is fixed and the directions of different bond vectors are completely uncorrelated. This corresponds to a three-dimensional random walk in space.

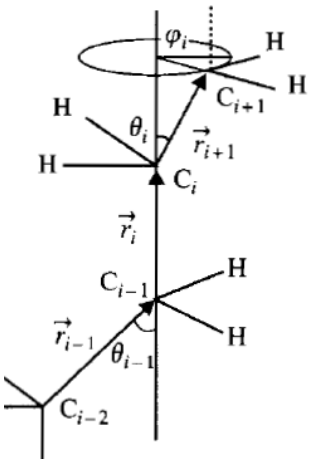


Figure 1: Three basic parameters in polymer models [4]. Bond length is the magnitude of the bond vectors. Bond angle θ is the angle between successive bond vectors. The zero value of the torsion angle ϕ corresponds to the bond vector r_{i+1} being collinear to r_{i-1}.

More complex polymer models account for non-uniform distributions of torsion or bond angles or correlations between bond vectors. These correlations could introduce considerable difficulty in mathematical treatment. For instance, the freely-rotating model assumes fixed bond angles, thereby introducing an additional parameter describing the decay rate of this local correlation; the mathematical expressions for properties (such as the mean-square end-to-end distance) in this model are also more complex compared with those of the random-flight model.

The difficulty of building an accurate polymer model for the coil segments in unfolded peptides is two-fold: the functions describing the torsion and bond angle distributions are presumably not in simple-forms; in addition, correlations between residues are not negligible. Moreover, these correlations include several nonlocal

components. One important nonlocal correlation is the excluded volume effect, which is simply the result of the fact that two different residues cannot occupy the same location at the same time. Therefore, the location of a given residue does not only depend on its immediate neighbors; it could depend on residues far apart in primary sequence. The nonlocal correlations add a significant layer of complexity to the polymer model because they abrogate the Markov property of the chain.

In summary, the polymer model for the coil segments should at least include:

(a) distribution of bond angles;

(b) distribution of torsion angles;

(c) excluded volume effects.

The distribution of bond and torsion angles can be inferred from available information on solved protein structures from PDB. Solution to the mathematical difficulty induced by the excluded volume effect, however, is not immediately apparent. One approach, of course, is to account for this effect by modifying equations from simpler models with empirical parameters [3]. This might allow us to fit for or even predict experimental observables quite well; however, it does not help us understand the problem unless we know the meaning of these parameters.

Another approach is to look for rigorous analytical expressions of the excluded volume effect, the expression for which will presumably be quite complex. Such degree of mathematical complexity digresses from the initial intent of constructing our model

(as explained in the main text), which is to build a minimally parameterized coarse-grained model for the unfolded proteins that is sufficient to explain the experimental data, but will enhance our understanding of important aspects of protein states and behavior.

1.4 Residue-to-residue distance distribution

Once a model is chosen, we need experimental data that well constrain the parameters of the model. In other words, the results of the chosen experiments should have connections to the HCPP model parameters. The method we chose is SAXS (small-angle X-ray scattering). In particular, we utilized a special type of SAXS (the anomalous SAXS) for our studies. As we can see in Chapter III and IV, these data can provide information regarding the residue-to-residue distance distribution of LRH1x, which is an important constraint to the HCPP model.

Next, I will briefly describe the underlying relationships that connect the HCPP model parameters with the anomalous SAXS curves. This theoretical bridge goes in two directions: on one hand, we need to know how to use the model to predict anomalous SAXS data; on the other hand, we should be able to use the anomalous SAXS data to train the model parameters.

The first step in data prediction is to obtain the ensemble composition using the HC theory, as outlined in the first section. For each ensemble member, we can calculate its residue-to-residue distance via the HCPP model. The distances within the helix

segments can be obtained via standard helix geometry; and those within the coil segments can be computed via the polymer physics model. After we have calculated the residue-to-residue distance for each ensemble member, we can use the Debye equation (described in Chapter III) to predict the corresponding anomalous SAXS signal.

Once we are familiar with the relationships that link HCPP model parameters with anomalous SAXS data, we can then use the data to train our model via a Bayesian network. Basically, a Bayesian network is a probabilistic graphic model that represents the conditional dependencies of a set of random variables. Our group is in the process of developing such a dynamic Bayesian network to model sequences of variables, for example, protein sequences.

Currently, Roy Hughes (my collaborator in Oas lab on the unfolded protein project) is working to create a theoretical map from a helix-coil ensemble to an residue-to-residue distance distribution. In parallel, he is devising a scheme which can parameterize such a mapping so that the parameters can be estimated from SAXS data. This mapping remains a long-term goal of the unfolded protein project.

2. Synthesis of Se-and Hg-labeled LRH1x

2.1 Structure and properties of LRH1x

In choosing a model polypeptide for studying the unfolded state, we must take into consideration the following aspects:

(a) The unfolded state should be considerably populated under experimental conditions.

(b) The peptide should have certain degree of secondary structure propensity and play an important role in the folding process of its parent sequence (if any). Also, the ensemble should change according to variation in certain experimental parameters so that we can use this to design a statistical mechanical model of the ensemble.

(c) Enumeration of the ensemble of conformations accessible to the polypeptide should allow the prediction of experimental observables.

(d) The peptide should be reasonably short. This will simplify the ensemble member enumeration and subsequent calculations while still providing proof-of-principle results. In addition, tertiary contacts will need to be accounted for in long peptides, which is beyond the current capability and goals of our HCPP model.

(e) The peptide should be soluble.

λ repressor is a protein found in phage lambda that binds to DNA with the helix-turn-helix motif (HTH) and regulates its transcription activity. In the self-assembled

dimer form [5], this protein binds with different affinity to different operator sites and keeps phage lambda in the lysogenic cycle.

The monomeric λ repressor consists of five helices with loops and turns to forms a single hydrophobic core [6] (Fig.2). The HTH begins with residue Gln33 and involves helix 2 to helix 3. Residues in helix 5 form the dimerization interface, and the N-terminal arm (first six residues) is critical for the operator site recognition and binding affinity [5].

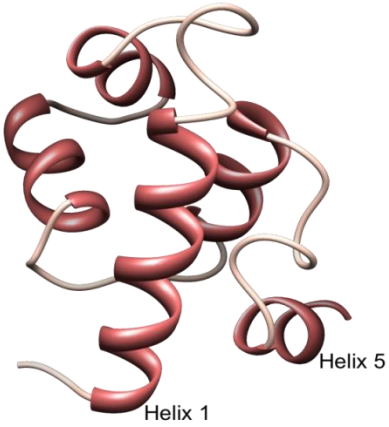


Figure 2: Structure of monomeric λ repressor.

An NMR study [6] on the N-terminal domain of monomeric λ repressor found that helix 1 has considerable intrinsic helicity, while the other regions are either more flexible or devoid of significant secondary structure. CD studies (by Jonathan Li in our lab) showed that the helicity of the peptide corresponding in sequence to helix 1 (LRH1x) changes with pH and temperature (Fig.3S). This gives us simple ways to modulate the ensemble composition of LRH1x by changing temperature or solution conditions. The peptide contains only 24 residues and it can be biosynthesized in E.Coli

via a TrpLE fusion protein expression protocol. Therefore, I chose LRH1x as the model peptide for investigating the unfolded state of proteins.

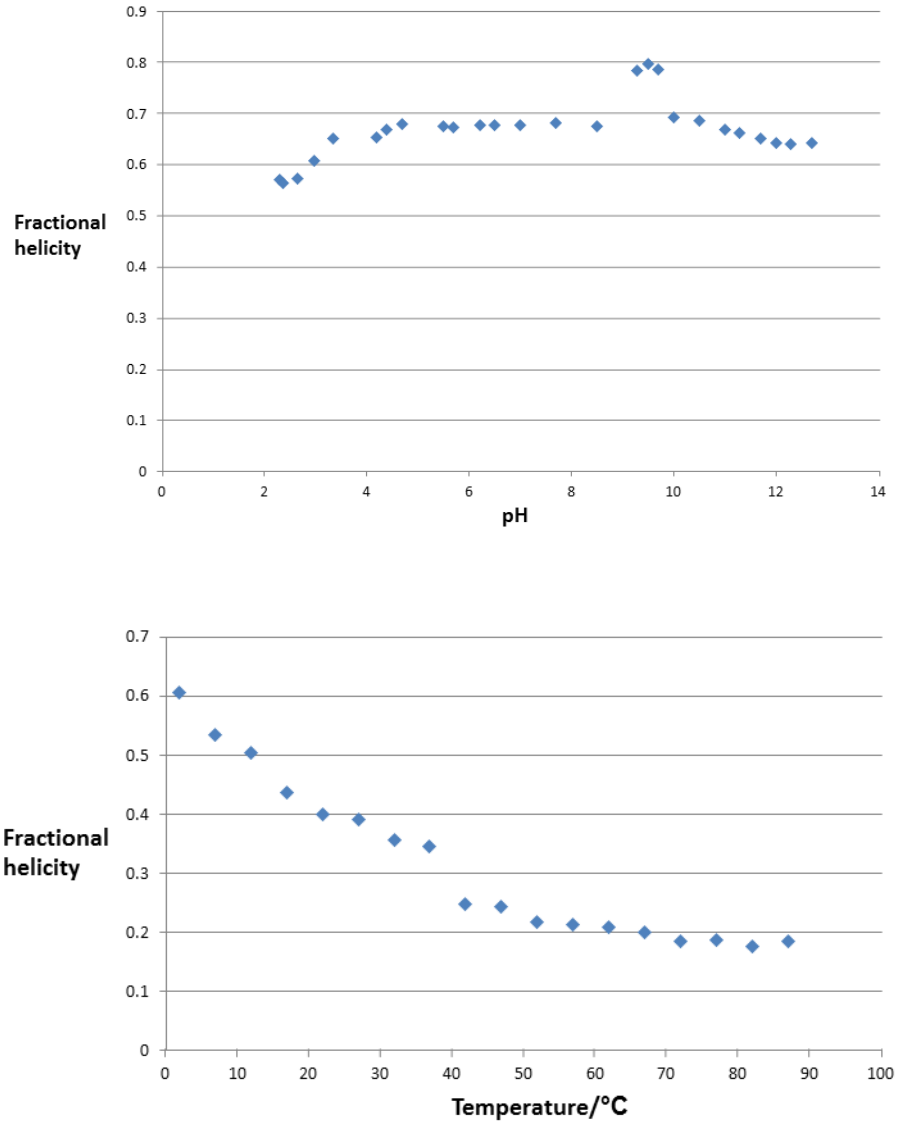


Figure 3: Average helicity of LRH1x changes with pH and temperature. Fractional helicity is calculated from CD signal by CDPro.

In order to label the LRH1x peptide with selenium or mercury and measure its end-to-end distance distribution, we inserted two cysteine residues at the N- and C-termini of the peptide. This modified version of LRH1x is denoted as LRH1xC.

2.2 Expression and purification of LRH1x

The expression of LRH1xC followed a TrpLE fusion protein preparation protocol, where the LRH1xC DNA sequence was joined with a leader sequence to form a fusion protein sequence. This fusion protein sequence was inserted into the plasmid of E.Coli, controlled by a T7 promoter. This fusion protein shows very low solubility in water. It easily forms inclusion bodies, therefore separating itself with other soluble proteins in the cell lysate during centrifugation, facilitating the purification process. Moreover, ten histidine residues were inserted into the leader sequence so we can separate it from LRH1xC after methionine cutting. This TrpLE expression system allows efficient biosynthesis of short peptides. It is less expensive but more accurate and environmental-friendly compared with the alternative chemical synthesis.

One liter of plasmid-transformed BL21 DE3 cells were grown to OD=0.85, when they were induced with 1 mM IPTG. The cells were incubated for another 4-5 h, after which they were harvested by centrifugation at 5000 rpm for 45 min and frozen at -78 °C.

After that, the cells were lysed and the fusion protein was separated from most other proteins by multiple steps of washing and centrifuging. To further purify the

fusion protein, a Ni-NTA column was used and the fusion protein was eluted at pH=5.0 in the middle of a pH elution gradient. The product was then dialyzed against 2% acetic acid to remove the salts introduced during the column step.

CNBr has been shown to react with methionine and cut proteins at these sites. Since the leader protein and LRH1x is connected with a single methionine and there is no methionine in LRH1x, we were able to separate these two parts with CNBr. During this reaction, the pure fusion protein was mixed with 1 g CNBr in 10 mL 70% formic acid. The reactant mixture was kept at room temperature for 1 h while spinning. Then the reaction flask is connected to a vacuum system so that the remaining liquid can be removed from the protein products. After 4-5 hours of degassing, the flask contains only a solid mixture consisting of LRH1xC, the leader and the unreacted fusion protein.

The leader and the fusion protein are separated from LRH1xC via another Ni-NTA column. Since the His-tag is present in the leader and the fusion protein but not in LRH1xC, LRH1xC will elute with the basic buffer while the other two proteins will remain on the column until they are washed off with acidic buffer. Then the LRH1xC was run through a G10 sizing column to remove the salts. After that, the peptide was run through a G25 sizing column to separate it from the remaining leader and fusion protein. Finally the LRH1xC peptide was lyophilized and stored at -78 °C for later use.

2.3 Selenium and mercury labeled LRH1x

In chapter III I describe ASAXS experiments on selenium and mercury labeled LRH1x. The peptide was labeled at the N- and C- termini and it was used for obtaining information on the end-to-end distance distribution.

2.3.1 Biosynthesis and purification of Se-LRH1x

The Se-LRH1x was biosynthesized using a cysteine-auxotrophic strain of E.Coli following the protocol outlined in S.Muller's 1994 paper [7] with certain conditions (selenocystine concentration, induction time, etc.) re-optimized for our particular system. This peptide was expressed with a TrpLE fusion protein system as explained in the last chapter. After cell harvest, the protein was purified through inclusion body preparation, nickel column #1, dialysis, CNBr cleavage, nickel column #2, G10 and G25 sizing column. This TrpLE expression system allows efficient biosynthesis of short peptides. It is less expensive but more accurate and environmental-friendly compared with the alternative chemical synthesis.

The expression of double-selenium labeled LRH1xC was first revealed by SDS-PAGE (Fig.1). A final yield of 5 mg was obtained from 1 L cell culture.

In order to verify the molecular weight of the product, ESI-MS was performed on purified double-selenium labeled LRH1xC (Fig.2a). The main peak (Fig.2b) has a molecular weight of 3116 Da, corresponding to the double-selenium labeled LRH1xC. As expected, peaks of the single-selenium labeled and unlabeled LRH1xC were

observed, whereas in much weaker intensity. A minor peak of 2960 Da was also present, which might result from the incorporation of two alanines when the cells were deprived of both cysteine and selenocystine during incubation.

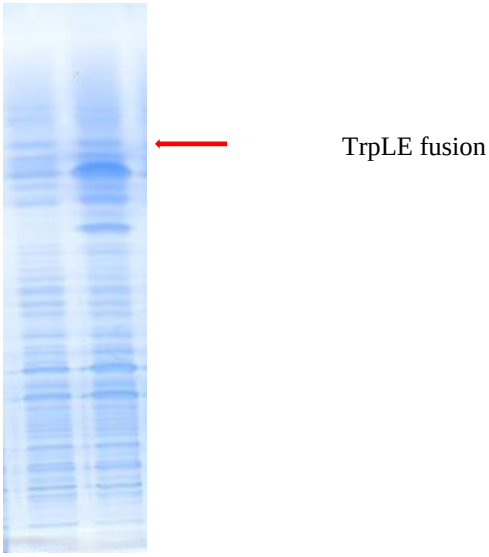


Figure 4: SDS-PAGE image of cell lysate before and after induction with IPTG. Left lane: before induction. Right lane: 6 h after induction.

2.3.2 The mercury labeling reaction

The data shown above established that the selenium label is incorporated into the peptide during expression. By comparison, the mercury label is not incorporated into the peptide until pure LRH1xC is obtained.

Thimerosal is a compound widely used as a preservative in vaccines and other drugs. In 2009, it was discovered that cysteines in proteins can be mercury-labeled by thimerosal in aqueous media. Through hydrolysis, thimerosal decomposes into

thiosalicylic acid and ethylmercury (EtHg), and the latter reacts with the sulfhydryl group in cysteine to form an adduct [8] (Fig.6).

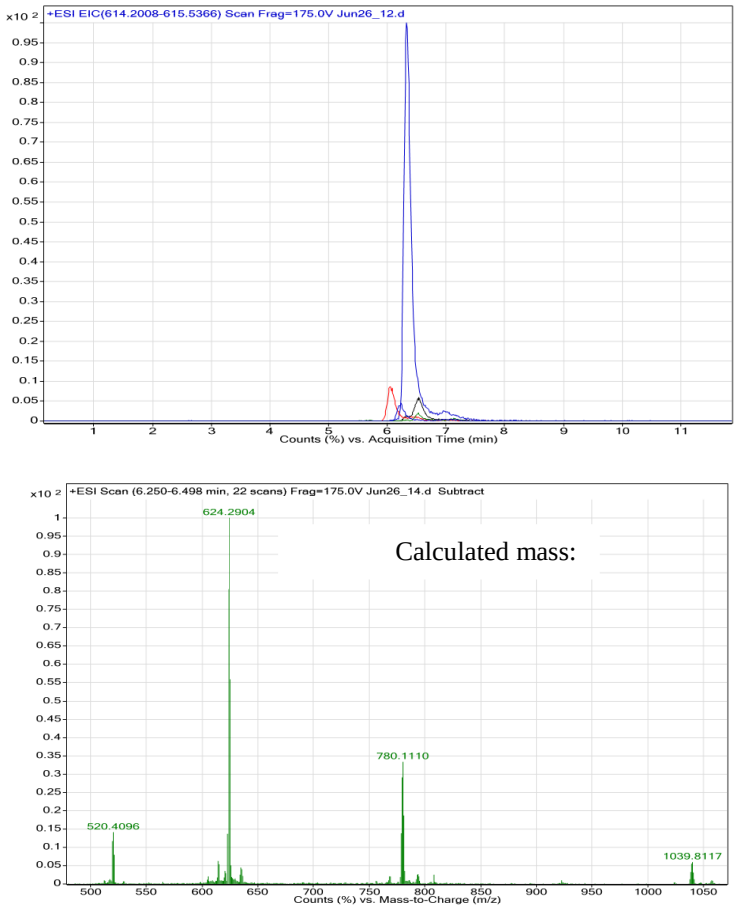


Figure 5: ESI-MS of Se-LRH1xC. The elution profile is shown in (a). Chromatogram for double selenium labeled LRH1xC is shown in (b).

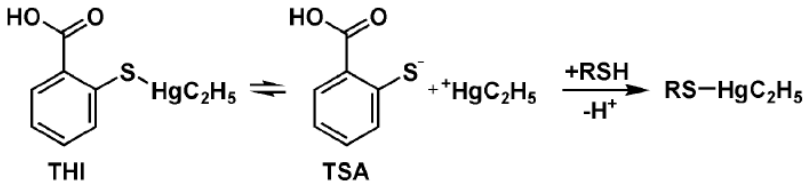


Figure 6: Hydrolysis of thimerosal in aqueous media and the formation of thiol-EtHg adduct.

Due to the high affinity of Hg to sulfur, the labeling reaction can be completed within 10 min in Tris-HCl buffer pH 8.5 (Fig.7). The lyophilized LRH1x was resuspended in deionized water and concentration was measured by the Edelhoch method. Then, stock solutions of 1 M Tris pH 8.0 and 5 M NaCl were added to a final concentration of 10 mM Tris, 100 mM NaCl and 35 μM. Next, an equal volume of 10 mM Tris, 100 mM NaCl and 1 M thimerosal was added to the above peptide solution. Incubate the reaction mixture at room temperature for about 10 min and buffer exchange the resultant solution into deionized water over a G10 sizing column to remove unreacted thimerosal. Finally, the peptide was lyophilized and ready for use. The final purity was assessed by MULDI, where only very small peaks of impurities show up. However, the peak intensity in MALDI is not directly related to the amount of a particular component in the sample. Due to instrumental limitations, we did not perform HPLC measurements to calculate the exact purity of the final product.

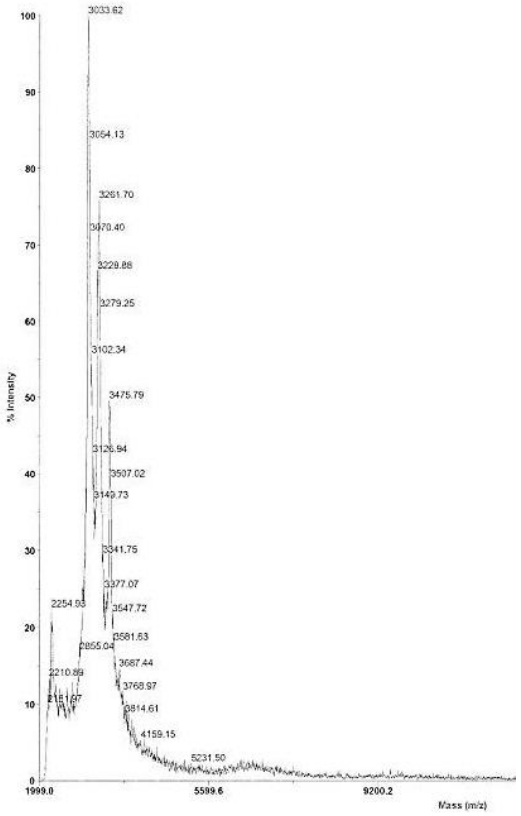


Figure 7: Mass spectrum result of the mercury labeling reaction. The peak at 3033.62 Da, 3261.70 Da and 3475.79 Da are unlabeled, single-Hg labeled and double-Hg labeled LRH1x respectively.

3. Small-angle X-ray scattering (SAXS)

3.1 Overview

SAXS is a technique used to determine the structure of proteins in terms of parameters such as shapes and sizes [9]. We can obtain such information by observing the way a distribution of macromolecules scatters at very low angles. To date, SAXS has proven to be an efficient way to determine low resolution protein structures due to its moderate requirement of protein sample preparation (compared with X-ray crystallography) and the ease of solution-phase measurements. However, all SAXS profiles are orientationally averaged because molecules tumble freely in solution. As a result, we can only obtain averaged structural information from scattering curves. In principle, a SAXS profile contains the pairwise distance information for all the atom pairs in a given protein, but reconstructing protein structures from SAXS curves often fails to provide a unique and stable solution due to the insufficient information content of an isotropically averaged SAXS profile. This leads to a considerable amount of uncertainty in the resulting structure; a "low-resolution" structure is obtained instead of an atomic level structure characteristic of crystallography and NMR methods.

In order to determine site-specific structural information from SAXS, I am developing an innovative anomalous SAXS method that allows accurate measurement of point-to-point distance distribution in a given protein. This technique will allow us to

directly observe the distance distribution between any two labeled amino acids in an unfolded protein.

Anomalous SAXS is a phenomenon observed in SAXS experiments where an atom (the label) absorbs strongly when the incoming X-ray is within a specific energy range (on-edge), giving rise to a sharp absorption peak in the edge scan; outside this narrow energy range (off-edge), the anomalous scattering becomes negligible compared with the normal SAXS scattering [10]. The anomalous scattering curve of a double-labeled protein can be converted into the label-to-label distance distribution via an indirect Fourier transform. This is the first use of anomalous SAXS for providing structural information about an unfolded protein.

3.2 Basic principles of SAXS

X-rays are photons whose wavelength ranges from 0.1 Å to 100 Å. At the Advanced Photon Source (Argonne National Laboratory, IL) where I performed my SAXS experiments, the X-ray is generated via a synchrotron source. The synchrotron radiation is a result of the acceleration of electrons through magnetic fields. These electrons are accelerated into the X-ray range, and the emitted X-rays are utilized to irradiate the sample. In most SAXS experiments, the X-rays directed at the sample are collimated and monochromatic, and the intensity of the scattered X-ray is measured by detector a positioned to detect scattered X-rays at very low angles incident to the incoming beam [9]. This scattered intensity is recorded at different scattering angles, and

the resultant I- θ function is the key data obtained from SAXS experiments. The following introduction is adopted from relevant chapters in Ref. [11].

In nature, X-ray scattering involves the acceleration of electrons by the Coulombic driving forces exerted by the X-rays. The electric field of the X-ray is oscillating, and the electrons are forced to oscillate at the same frequency as the incident radiation. Since an accelerated charge emits electromagnetic radiation, this oscillating electron then becomes a new source of X-rays whose frequency is the same as the incident radiation. It is through this process that the electrons scatter the incident X-rays in all directions.

Mathematically, the propagation of X-ray in the **k** direction can be expressed by a complex function (vectors are represented in bold):

$$E(\boldsymbol{r},t) = E_0 exp\left[2\pi i\left(\frac{kr}{\lambda} - \upsilon t + \delta\right)\right]$$

$$= E_0\left\{cos\left[2\pi i\left(2\pi i\left(\frac{kr}{\lambda} - \upsilon t + \delta\right)\right)\right] + isin\left[2\pi i\left(2\pi i\left(\frac{kr}{\lambda} - \upsilon t + \delta\right)\right)\right]\right\}$$

where E(**r**,t) is the electric field at point **r** and time t; **k** is a unit vector in the **k** direction; λ is the wavelength; ν is the frequency; δ is the phase of the wave (in cycles) and E_0 is the maximal amplitude. This equation is known as the Euler's formula

These two forms of Euler's formula are equivalent, but for mathematical convenience, we choose the first expression (the exponential form) for the following analysis.

The basic scheme of a SAXS experiment is shown in Fig. 8. A beam of collimated X-rays shines on the sample. The scattered X-rays at various scattering angles are recorded by the detector, while the unscattered X-ray is absorbed by a beam stop. $\mathbf{s}_0$ and **s** are unit vectors (equivalent to k in Euler's formula) that represent respectively the incoming radiation and the scattered wave. The scattering angle is defined as half of the angle between these two vectors. The basic geometry of this setup is shown in Fig. 9.

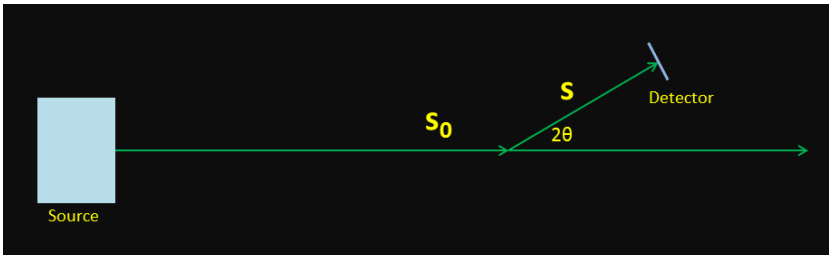


Figure 8: Basic scheme of SAXS experiments.

The **S** in Fig.8 is usually defined as the scattering vector:

$$\mathbf{S} = (\mathbf{s}/\lambda) - (\mathbf{s}_0/\lambda)$$

The dimension of **S** is reciprocal length (since s is unitless and λ has the unit of length) and the magnitude of this vector can be calculated as:

$$|\mathbf{S}| = (\mathbf{S}\cdot\mathbf{S})^{1/2}$$

$$= [(\mathbf{s}^2+\mathbf{s}_0^2-2\mathbf{s}\cdot\mathbf{s}_0)/\lambda^2]^{1/2}$$

$$= [(1+1-2\cos 2\theta)/\lambda^2]^{1/2}$$

$$= (4\sin^2\theta/\lambda^2)^{1/2}$$

$$= 2\sin\theta/\lambda$$

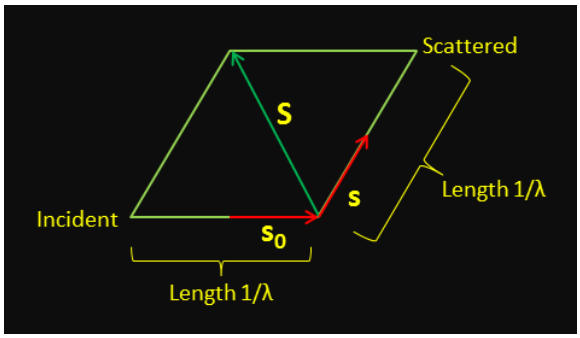


Figure 9: Basic geometry of SAXS experiments.

For a single electron at the origin **O** (Fig.10), the radiation E(**S**) seen by the detector can be computed by quantum mechanics. However, a real sample usually contains many electrons, and most of them are not located at the origin. In order to calculate the radiation scattered by a real sample, we introduce a new parameter called the structure factor, F(**S**). It is defined as the ratio of the radiation scattered by the real sample to that scattered by a single electron at the origin. Fig.9 shows the case where a second electron is located at position **r** relative to one at the origin. Because moving the electron from the origin to position **r** caused a phase shift of **S**·**r** cycles, the scattered radiation by the second electron is then E(**S**)exp(2πi**S**·**r**). The structure factor for the second electron is therefore exp(2πi**S**·**r**).

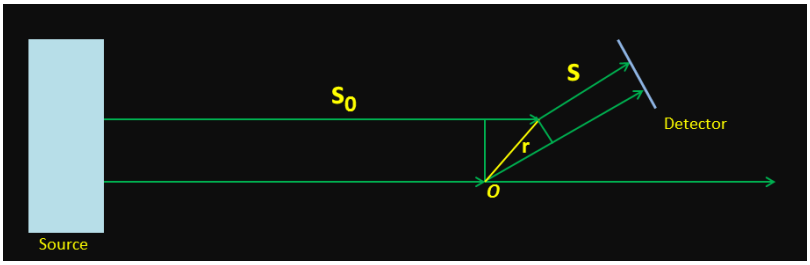


Figure 10: X-ray scattered by electron located at position r.

A better way to describe the continuous distribution of electrons in a real sample is using the electron density $\varrho(\mathbf{r})$ in a volume element $d\mathbf{r}$ located at $\mathbf{r}$. The structure factor for a real sample can then be expressed as the integration over the entire sample:

$$F(\mathbf{S}) = \int d\mathbf{r}\, \varrho(\mathbf{r}) \cdot \exp(2\pi i \mathbf{S}\cdot\mathbf{r})$$

From the above expression, we can see that if the electron density distribution of a sample is known, the structure factor and the corresponding X-ray scattering can be computed.

We can also see from this equation that the structure factor is actually a Fourier transform of the object. Consequently, the electron density distribution is the inverse Fourier transform of the structure factor.

However, the structure factor $F(\mathbf{S})$ is not directly measurable in SAXS experiments. What we actually observe is the intensity of scattered X-rays at different scattering angles. The intensity is a real number expressed as

$$I(\mathbf{S}) = F(\mathbf{S})F^*(\mathbf{S}) = |F|^2$$

The Fourier transform of intensity is the pairwise distance distribution function (PDDF):

$$PDDF = \int d\mathbf{r} \varrho(\mathbf{r}) \varrho(\mathbf{u}+\mathbf{r}) = \int d\mathbf{S}\, I(\mathbf{S}) \exp(-2\pi i \mathbf{S}\cdot\mathbf{r})$$

where $\mathbf{u}$ can take any value that $\mathbf{r}$ can. The PDDF is essentially a map of all inter-scatterer distances within the sample. As a consequence, if we could obtain $I(\mathbf{S})$ (or I(q) as commonly used in the SAXS community, where $q=2\pi|\mathbf{S}| = 4\pi|\sin\theta|/\lambda$),

information regarding the structure of the sample can then be inferred according to the above equation.

From the above analysis, we begin to see why SAXS is considered as a type of scattering. Scattering is commonly defined as the process where a radiation deviates from a straight trajectory caused by heterogeneity of the medium. In the case of SAXS, if the sample is homogeneous, then the electron density $\varrho(\mathbf{r})$ follows a uniform distribution. Therefore,

$$F(\mathbf{S}) = \varrho \int d\mathbf{r} \exp(2\pi i \mathbf{S}\cdot\mathbf{r})$$

This is just the Dirac delta function $\delta(\mathbf{S}-0)$, meaning the only X-rays that emerge from the sample are F(0), which is parallel to the incident beam. In other words, a homogeneous sample cannot scatter X-rays at all. In order for scattering events to occur, there needs to be a contrast in $\varrho(\mathbf{r})$ between a given region and its neighbors. This is the basic requirement for SAXS, as for all other scattering phenomena.

3.3 Basic principles of anomalous SAXS (ASAXS)

Generally, when a beam of X-rays hits the sample, the oscillating electric field of the electromagnetic radiation interacts with the electrons bound in an atom [12]. The radiation may undergo normal scattering as described in the preceding section, or it can be absorbed and excite the electrons to produce photoelectrons, as shown in Fig.11.

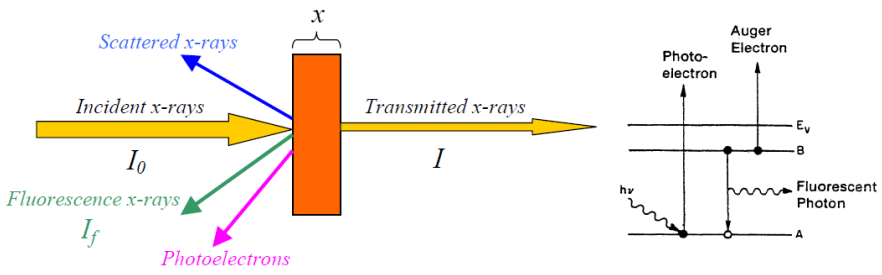


Figure 11: Interactions between incident X-rays and the sample [12].

If we shine a narrow beam of monochromatic X-rays upon a sample, the intensity of transmitted X-rays follows the expression:

$$\ln (I_0/I) = \mu t \varrho$$

where μ is the linear absorption coefficient depending on the type of the material, ϱ is the material density, t is the sample thickness, I_0 and I represent the intensity of the incident and the transmitted beam respectively. We usually define this decrease in intensity observed in transmitted X-rays as X-ray absorption. By fixing I_0 and measuring I at different incident X-ray energies (E_{in}), we can determine the intensity of absorbed X-rays (I_{abs}). This I_{abs} - E_{in} plot is referred to as the X-ray absorption spectrum.

One striking feature of these X-ray absorption spectra is that they are not smooth curves, but exhibit step-like features: at certain energies the absorption increases drastically, giving rise to an absorption edge (Fig.12 shows an example). The occurrence of such absorption edge leads to anomalous X-ray scattering. Each edge occurs when the energy of the incident photons is just sufficient to excite a core electron of the absorbing atom to a continuum state, i.e., to produce a photoelectron. Thus, the energies of the

absorbed radiation at these edges correspond to the binding energies of electrons in the K, L, M…shells of the absorbing elements. The edges are consequently named in this manner: for instance, K-edge is the sudden increase in absorption resulted from the excitation of a K-shell electron. The energy of absorption edges are dependent on the type and valence state of the element as well as the environment it sits in.

The anomalous scattering phenomenon can be understood as the resonance between the incident X-rays and the electronic transitions from bound atomic orbitals to electronic states in the continuum [10]. The quantum-mechanical derivation of the anomalous scattering factors is quite complicated, but a classical view where bound electrons are treated as dampened oscillators can also reveal qualitative features of this phenomenon.

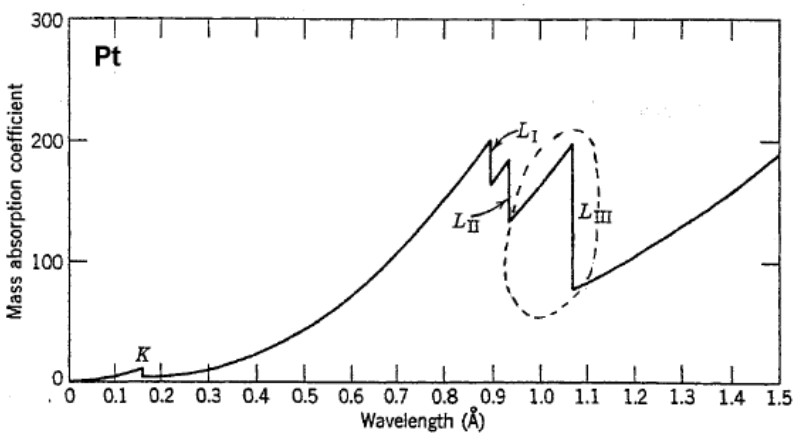


Figure 12: X-ray absorption spectrum of Pt [9].

Mathematically, the oscillation amplitude of a bound atomic electron is expressed as

$$x_0 = -\frac{eE_0}{m\left(\omega^2 - \omega_0^2\right)}$$

where E_0 and ω are the maximal amplitude and the frequency of the incident beam respectively, e and m are the charge and mass of the electron respectively, ω_0 is the natural vibration frequency of the dipole oscillator corresponding to the net binding force exerted on the electron by the nucleus as screened by other electrons. This amplitude becomes very large when ω is close to ω_0, giving rise to the edges on the spectra, i.e., the anomalous scattering phenomenon. However, x_0 does not go to infinity when $\omega = \omega_0$ because there is a small damping effect due to the energy loss to classical radiation by the oscillating electron. Therefore, the electron's equation of motion in the damped, driven harmonic oscillator is

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -k'\frac{\mathrm{d}x}{\mathrm{d}t} - kx + eE_0\,\mathrm{e}^{i(\omega t - \delta)}$$

where the damping force is assumed to be proportional to the electron's velocity (F = -k'v). We can obtain the amplitude of scattered radiation (ε) by this bound electron via solving this equation. Dividing ε by that of a free electron (ε_0), we now have the expression for the scattering factor

$$f_e = \frac{\omega^2}{\omega^2 - \omega_0^2 - i\gamma\omega}$$

where

$$\gamma = \frac{2}{3}\left(\frac{e^2}{mc^2}\right)\frac{\omega^2}{c}$$

where c is the speed of light in vacuum.

One of the most important features of anomalous scattering is that the anomalous scattering factor is a complex number. In normal scattering, γ is much smaller than ω, so the term $i\gamma\omega$ can be neglected when calculating the normal scattering factor f_0, making f_0 a real number. However, in anomalous scattering where $\omega \approx \omega_0$, this imaginary term can no longer be neglected, giving rise to a complex scattering factor f_e.

$$\begin{cases} f_e = f_e' + if_e'' = \dfrac{\omega^2\left(\omega^2-\omega_0^2\right)+i\gamma\omega^3}{\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2} \approx 1+\left[\dfrac{\omega_0^2\left(\omega^2-\omega_0^2\right)+i\gamma\omega\omega_0^2}{\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2}\right] \\ \begin{cases} f_e' = \dfrac{\omega^2\left(\omega^2-\omega_0^2\right)}{\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2} \approx 1+\left[\dfrac{\omega_0^2\left(\omega^2-\omega_0^2\right)}{\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2}\right] \\ f_e'' = \dfrac{\gamma\omega^3}{\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2} \approx \dfrac{\gamma\omega\omega_0^2}{\left(\omega^2-\omega_0^2\right)^2+\gamma^2\omega^2} \end{cases} \end{cases}$$

Therefore, the total scattering factor f at the edge contains wavelength-dependent contributions f_e' and f_e'' in addition to the wavelength-independent normal scattering factor f_0. The expression for the total scattering factor is expressed as

$$f = f_0 + f_e' + if_e''$$

While the normal scattering factor decreases with increasing scattering angle, the anomalous scattering factor is virtually independent of scattering angle [10].

At the end of this section, I'd like to explain the word "anomalous" under this context. The reason why this phenomenon is often referred to as "anomalous" scattering is because it is not normally observed. The frequency of the incident X-ray has to be tuned to the right range (the edge) for the anomalous scattering to occur. However, this process per se is nothing anomalous; it is a natural process fairly well understood and is applicable to many studies.

3.4 The "Three-Energy Strategy"

In order to obtain the pure anomalous scattering signal corresponding to the residue-to-residue distance distribution, I labeled the two residues in question with a new element. The product of this labeling process is a doubly-labeled protein.

We have quite a few choices of the "new element". In general, its anomalous scattering edge should be distinct from that of the elements present in the protein (C, N, O, etc). Also, a large anomalous scattering factor is preferred so that the signal-to-noise ratio is acceptable. In my experiments, I used selenium and mercury as the two ASAXS labels of choice, details of the labeling process can be found in Chapter II.

For the ASAXS studies, I used a protein fragment from λ repressor (LRH1x). This intrinsically unfolded peptide is an ideal system with which to develop the ASAXS method because the average helicity of the peptide changes according to solution conditions. Chapter II has discussed the structure and properties of this peptide in more detail.

To extract the ASAXS signal from the double-labeled peptide, I followed the Three-Energy Strategy developed by G.Goerick et al [13]. where the authors used the ASAXS method to investigate the distribution of metal counterions around polymer chains. In this method, the SAXS curve of the chain is measured at three distinct energies, and the ASAXS signal was obtained by the following equation:

$$\begin{aligned} S_{\mathrm{Ion}}(q) &= 4\pi \int_{V_p} \int v(\vec{r}) v(\vec{r}\,') \frac{\sin\left(q|\vec{r}-\vec{r}\,'|\right)}{q|\vec{r}-\vec{r}\,'|} \mathrm{d}^3 r \, \mathrm{d}^3 r' = \\ &= \left[\frac{\Delta I_0(q, E_1, E_2)}{f'_{\mathrm{Ion}}(E_1) - f'_{\mathrm{Ion}}(E_2)} - \frac{\Delta I_0(q, E_1, E_3)}{f'_{\mathrm{Ion}}(E_1) - f'_{\mathrm{Ion}}(E_3)} \right] \cdot \frac{1}{F(E_1, E_2, E_3)}, \\ F(E_1, E_2, E_3) &= f'_{\mathrm{Ion}}(E_2) - f'_{\mathrm{Ion}}(E_3) + \frac{f''^2_{\mathrm{Ion}}(E_1) - f''^2_{\mathrm{Ion}}(E_2)}{f'_{\mathrm{Ion}}(E_1) - f'_{\mathrm{Ion}}(E_2)} - \frac{f''^2_{\mathrm{Ion}}(E_1) - f''^2_{\mathrm{Ion}}(E_3)}{f'_{\mathrm{Ion}}(E_1) - f'_{\mathrm{Ion}}(E_3)} \end{aligned}$$

The first equation is the expression of the pure anomalous signal, where the f' and f'' are the anomalous scattering factors and ΔI is the difference SAXS curve between the two energies indicated in the bracket. This method applies well to our case, where the polymer is our peptide and the metal counterions are the two ASAXS labels at the N- and C-terminal.

The decomposition of the anomalous SAXS profile is based on the Debye equation [9].

$$I(q) = \sum_{D=1}^{D_{max}} P(D) * f_A * f_B * \frac{sin(qD)}{qD}$$

where I is the scattering intensity, q equals $4\pi\sin(\theta)/\lambda$, D is the label-to-label distance, f_A and f_B represents the form factor of the two labels (in our case A=B), and P(D) is the relative population of the peptides with a label-to-label distance of D.

Therefore, the residue-to-residue distance distribution of the peptide could be determined in the following steps:

(1) Label the two sites with proper ASAXS labels;

(2) Measure the SAXS curve of this labeled peptide at three different energies;

(3) Use the Three-Energy Strategy to extract the pure anomalous signal;

(4) Fit the pure anomalous signal to a PDDF of the ensemble.

3.5 SAXS data collection of Se labeled LRH1x

SAXS data were collected at the Advanced Photon Source (APS) at Argonne National Laboratory outside Chicago. Sample concentration was approximately 5 mg/mL in a buffer solution consisting of 1.0 mol/L HEPES, 2.0 mol/L NaCl and 0.4 mol/L TCEP. Temperature was set to be 278 K throughout our experiment. Edge scan was performed on powder L-selenocystine to determine the edge energy for selenium in this particular chemical environment. Radiation damage and sample aggregation were carefully examined throughout our measurements.

By performing an edge scan with powder selenocystine (Fig. 13), the on-edge energy of the selenium anomalous scattering was determined to be 12654 eV. The three energies we chose were 12654 eV, 12651 eV and 12604 eV.

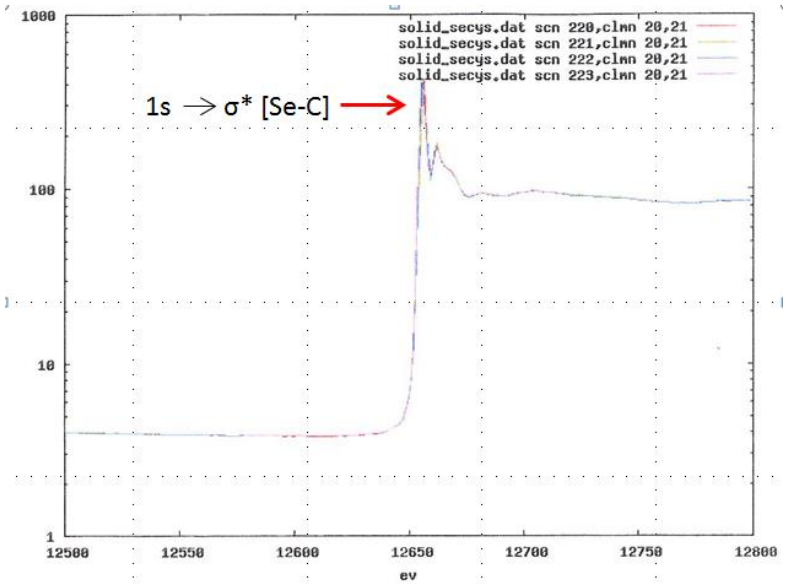


Figure 13: Edge scan of powder selenocystine. X-axis is the energy of incoming X-ray in eV. Y-axis is the absorption of X-ray in arbitrary units.

3.6 SAXS data collection of Hg labeled LRH1x

The basic data collection procedure was the same as that for selenium-labeled LRH1x. The buffer composition was 1% glycerol, 100 mM NaCl and 20 mM HEPES pH 7.4. Temperature was set to be 283 K throughout our experiment. Edge scan was performed on powder thimerosal to determine the L-III edge energy for mercury in this particular chemical environment. The edge was measured to be 12282 eV and the three energies for ASAXS measurements were 12280 eV, 12180 eV and 12080 eV. Two peptide concentrations were used in these measurements: 2 mg/mL and 10 mg/mL.

3.7 Fractal dimension of LRH1x

From these scattering curves, we can obtain some interesting information of the peptide without doing much data analysis, for example, its fractal dimension.

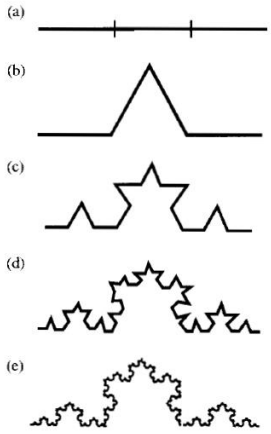


Figure 14: The Koch Curve.

Fractal structures are characterized by self-similarity within some spatial range[14]. In this case, the structure of the object is independent of the characteristic length scale of observation[14] . For regular structures, self-similarity is generally geometric. One famous example in this category is the Koch Curve (Fig. 14). For random or irregular objects, the observed self-similarity is statistical in nature. Unfolded proteins belong to the latter category. However, self-similarity in proteins only occurs within a limited range of scales. In this range, we can characterize the scaling property of the protein using a parameter called the fractal dimension.

As shown in Fig.15, the mass m enclosed by the sphere increases with the radius r according to a power-law relationship:

$$m \propto r^D$$

where D is defined as the fractal dimension (FD) of the object. A one-dimensional object (a line) has an FD of 1; a two-dimensional object (a plane) has an FD of 2; a three-

dimensional object (a solid) has an FD of 3. However, many objects have FD less than the spatial dimensions they occupy. In these cases, D is not an integer. For instance, a stiff rod has an FD of 1, a segment of helix has an FD around 2, and a well-solvated polymer has an FD of 1.7 with excluded volume effects accounted for. The FD for a random-flight chain that can cross itself (the ideal random-flight chain) is 2.0[15].

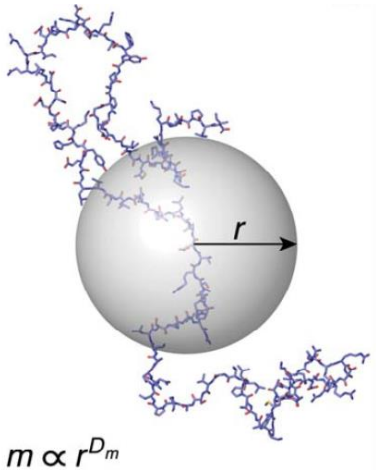


Figure 15: Fractal dimension in proteins[14] .

The FD for proteins can be measured by SAXS experiments. The scattering intensity changes with S (Italics indicate the norm of a vector) via the following equation:

$$I \propto q^{-D}$$

which gives rise to a linear region in the log(I) versus log(q) plot[15] ; the negative slop on this plot gives us the fractal dimension D. The determination of FD for LRH1x is shown in Fig. 16, where the FD is 1.41. This value seems to indicate that some segments

within the ensemble appear as stiff rods, bringing the observed FD lower than 1.7 expected for a well-solvated coil.

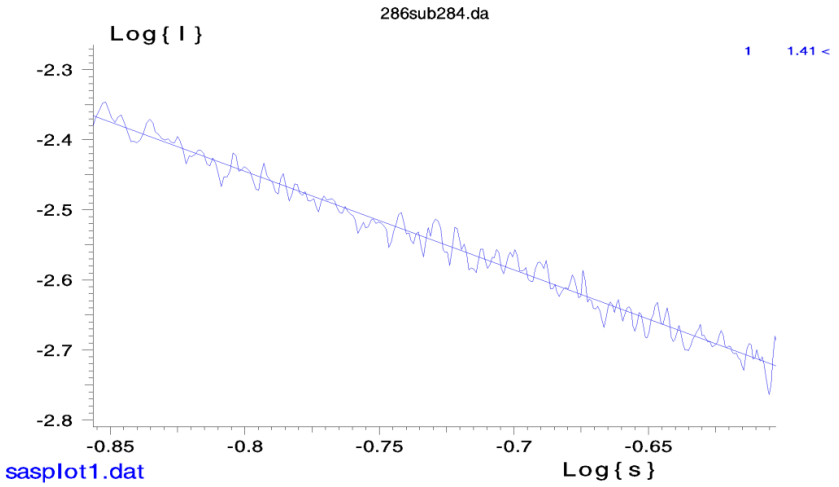


Figure 16: Fractal dimension of double Se labeled LRH1x. Log(s) on the x-axis is equivalent to log (q).

3.8 Overview of the selenium label

Among the many potential ASAXS labels, the selenium label (in the form of selenocysteine, or Sec) has its unique advantages. First of all, the selenocysteine can be incorporated into the LRH1x peptide via biosynthesis, as will be shown later. No subsequent labeling reactions are required once the peptide is successfully expressed and purified, which eliminates the need for another round of purification after peptide preparation. Another nice feature about the selenium label is its high similarity to sulfur so that the anomalous label does not introduce much perturbation to the unfolded ensemble. Se and S are both VIA elements in adjacent rows, leading to their similar chemical properties. Two important differences between selenocysteine and cysteine,

however, are their different pKa values and redox potentials. Selenocysteine has a lower pKa (5.5) than cysteine (8.3), and it is more reductive than cysteine. The former property might lead to minor perturbations to the ensemble distribution due to electrostatic interactions, which has yet to be determined by CD spectra (or the HCPP model can be devised to incorporate selenocysteine as the 21st amino acid); the latter prompts me to introduce more reducing agents (TCEP or DTT) into the peptide solution both in the purification process and at the beamline.

However, one major problem with the selenium label is its low yield during expression and unsatisfactory SNR in SAXS experiments. Usually, 5 mg peptide is expected from 1 L synthetic-rich media supplemented with selenocystine, making the SAXS sample preparation a costly process.

I was able to obtain the SAXS signal for the double-Sec-labeled LRH1x at three different energies. Using the Three-Energy Strategy explained earlier, end-to-end distance distribution was obtained, which is quite distinct from the expected.

3.9 End-to-end distance distribution of Se labeled LRH1x

By using the Three-Energy Strategy, the pure anomalous signal is obtained from the SAXS curves measured at the three distinct energies (Fig.17).

This provides us the information necessary to solve for the end-to-end distance distribution of the peptide using the Debye equation. However, due to the low

anomalous scattering intensity of the Se label and a corresponding poor signal-to-noise ratio (SNR), I failed to arrive at a unique solution of the label-to-label distance distribution. A low SNR decreases the information content of the SAXS profile. The number of equations available for solving the Debye equation is therefore inadequate. In such situations, one of the choices is to apply regularization methods or maximum entropy models that require prior information such as the maximum label-to-label distance. Here, I used the GNOM method [15] for this purpose. This method basically adds a smoothness constraint to the solution as the regularization parameter besides the Debye equation, decomposing the anomalous signal into a unique distance distribution. The maximum distance is a user-input and is set to be 82 Å in this case for a fully extended peptide chain. The end-to-end distance distribution obtained in this way is shown in Fig.18.

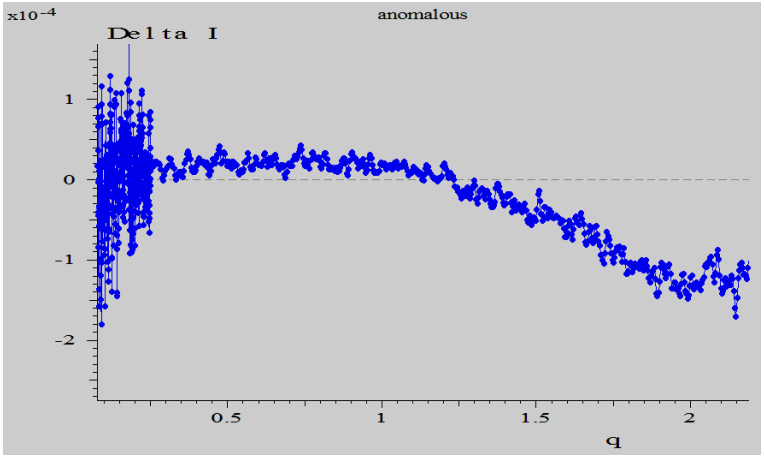


Figure 17: Pure anomalous signal of double Se labeled LRH1x.

First of all, we can clearly see that it is a broad distribution. However, this can also be an artifact of the regularization method we chose, since smooth curves are preferred over rough ones. But we can tell the distribution is likely to be broad from the anomalous curve, because this curve should show oscillating features (Fig.19) if only a few distances are present in the ensemble judging from the Debye equitation.

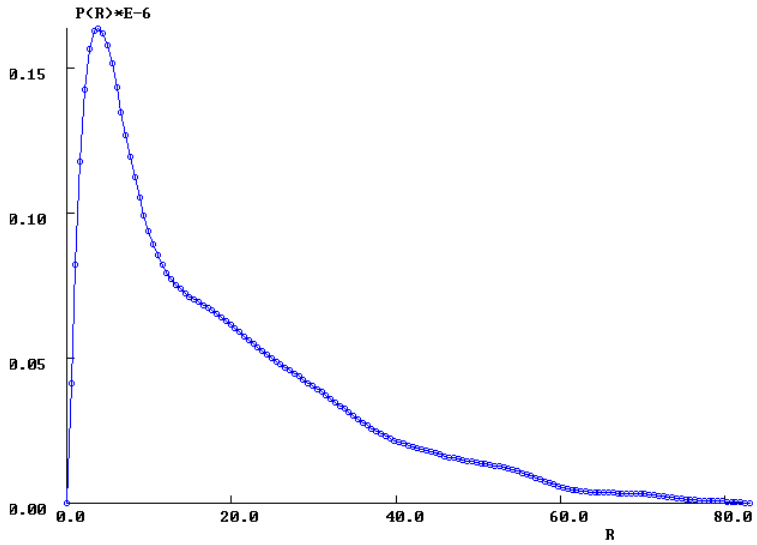


Figure 18: End-to-end distance distribution of double Se labeled LRH1x. X-axis is distance in angstroms and Y-axis is the relative probability of each distance.

The peak of the distribution is at 8 Å, which is rather unexpected. Some of the conformers are not very helical and can indeed guarantee a small end-to-end distance (Fig. 20), but whether these conformers constitute a major part of the whole ensemble remains an open question. The presence of this peak might be a result of the regularization method or it could have been caused by the low SNR. It could also a

consequence of limited amount of selenium labeled peptide samples since we were not able to repeat the SAXS measurement on site to ensure the quality of the data.

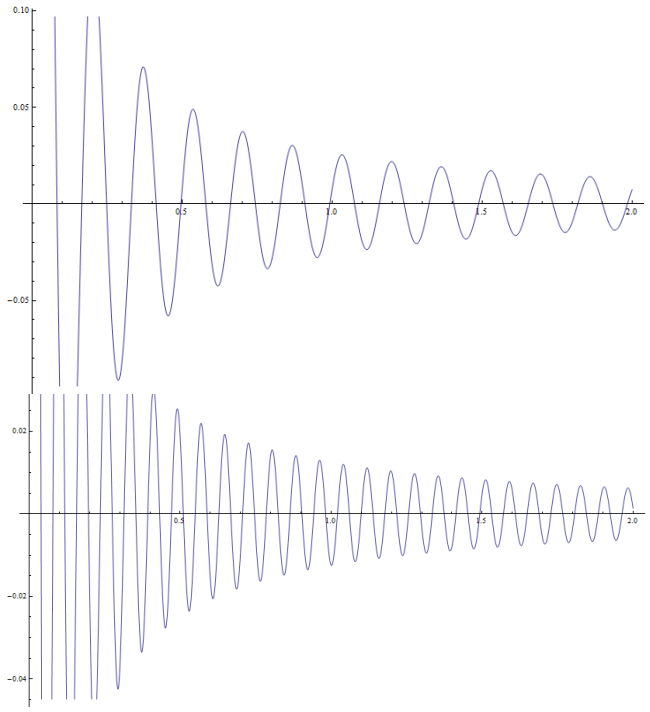


Figure 19: Simulated anomalous signal showing oscillating features. (Above) Anomalous signal from an ensemble all of whose members have an end-to-end distance of 38 Å. (Below) Anomalous signal from an ensemble all of whose members have an end-to-end distance of 80 Å.

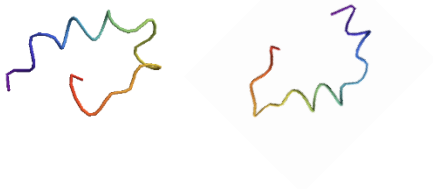


Figure 20: Two conformers in the LRH1x ensemble whose end-to-end distance is small compared with the maximum possible distance.

In order to increase the SNR of these experiments and improve the yield of the labeling process, we decided to try mercury as a new label in hope of better results.

3.10 Advantages and disadvantages of the mercury label

In considering using mercury as our new ASAXS label, we realized that it has several advantages when applied to our case. First of all, since we will not introduce the toxic selenium into the bacteria culture during their growth, the yield of the peptide is higher. And we will not need to add large amounts of reducing agents during peptide purification to protect the highly reductive –SeH group. Secondly, mercury shows high affinity and specificity to sulfur and this reaction goes to completion within one hour at room temperature. Thirdly, at the L-III edge, Hg has large anomalous scattering factors: its f'' is 10 compared with selenium's 3.8 at its K edge. Large scattering factors mean strong scattering, which could lead to better SNR of the resultant SAXS curves.

However, one major drawback of the mercury label is that it's much heavier and bigger than selenium. Also, it is not similar to sulfur or any other elements already present in the peptide, so it might perturb the original peptide ensemble. The effect of Hg labeling on helicity could be detected by CD spectroscopy.

3.11 ASAXS curves of the mercury labeled LRH1x

By using the Three-Energy Strategy again, we were able to obtain the ASAXS signal for mercury labeled LRH1x at 2 mg/mL and 10 mg/mL (Fig.21). As we can see, the anomalous scattering is quite weak and the intensity is essentially around zero. Although we did not observe peptide aggregation or radiation damage in either of the

two samples, the ASAXS curves obtained from the two concentrations show very different features and are both noisy. We conclude that the difference we observe in these two curves here is not due to real differences of our samples; rather, it is a result of the poor SNR in our experiments.

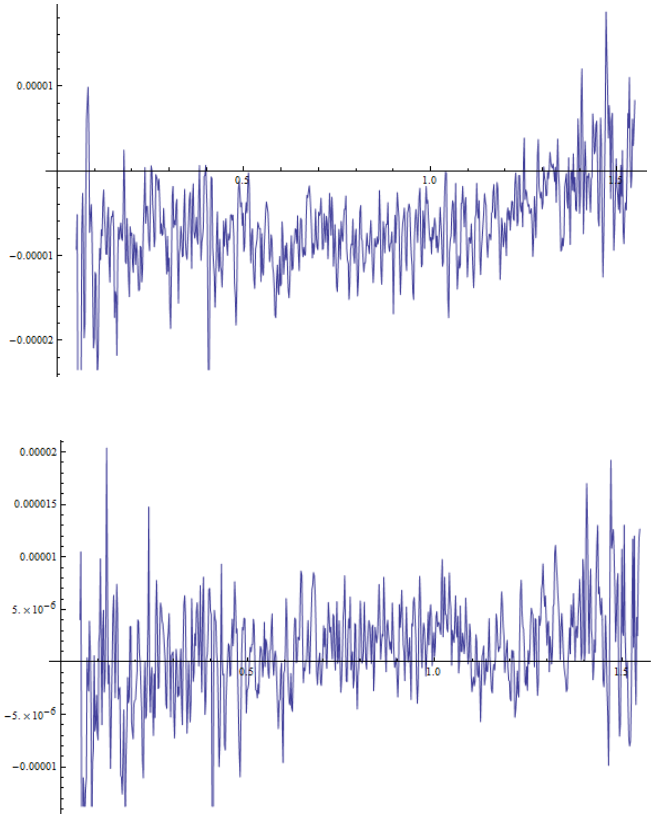


Figure 21: Pure anomalous signal of double Hg labeled LRH1x. (Above) 2mg/mL. (Below) 10 mg/mL. X-asix is q in $Å^{-1}$. Y-axis is the anomalous scattering intensity in arbitrary units.

We decide not to continue with the ASAXS signal decomposition with this noisy data. One piece of good news at the beamline though is that aggregation didn't occur at 10 mg/mL LRH1x, which was the maximum concentration available due to the limited

amount of labeled peptides. Therefore, one possible solution to the low SNR is to increase sample concentration during measurements.

However, before modifying the current experimental procedures with selenium or mercury, we need to make sure that these ASAXS curves give us useful information about the peptide ensemble. To this end, I carried out a series of simulations in Mathematica using the Three-Energy Strategy in order to find out whether the ASAXS curve is sensitive to ensemble composition. Chapter IV offers a detailed description of these simulations and conclusions are made at the end of the chapter regarding the viability of these ASAXS experiments.

4. Simulation of the ASAXS signal

In order to determine the feasibility of the ASAXS experiments, simulation is necessary since it provides information about the sensitivity of ASAXS curves to ensemble compositions. If the shape and features of an ASAXS curve hardly changes upon changing of ensemble compositions, then we conclude that the ASAXS curve does not contain much information of the ensemble, in which case we will not continue experimenting with this method. On the other hand, if the ASAXS curve changes according to the input ensemble composition, then we say it truly contains information of the ensemble and it is worthwhile to improve the SAXS method itself or the labeling technique.

Because we do not have a mature HCPP model at this stage, we are not yet able to generate correct estimations of the ensemble composition at any given experimental condition. However, we can simulate the end-to-end distance distribution of an unfolded peptide ensemble as a Gaussian distribution. First, I ran the simulation with single Gaussian distributions which mimics a highly helical ensemble. To simulate a more realistic, more helical system, I then simulate ASAXS curves using distance distributions resulted from a combination of many different Gaussian distributions. The form factors I used in these simulations are those of mercury, since it is more likely to give us good SNR compared with selenium as discussed in the last chapter. By comparing the SNR from previous ASAXS experiments and the intensity difference

among these simulated curves, we can know whether the anomalous signal is significant enough for decomposition.

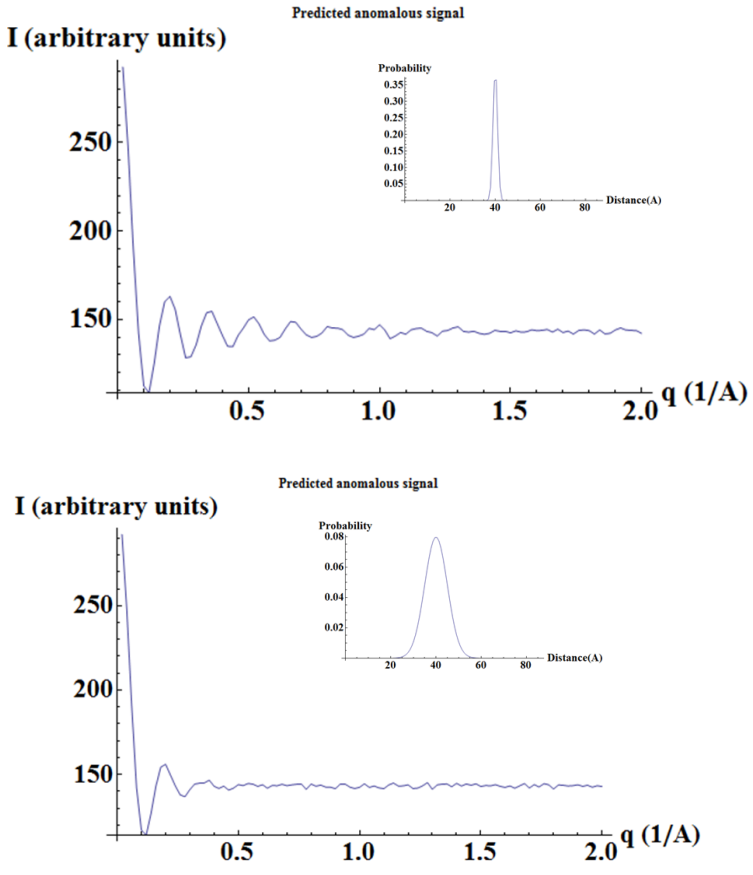
4.1 Simulation of ASAXS signal using single Gaussian distance distribution

As an initial test of the method, I simulated the ASAXS signal from the double labeled peptide using a single Gaussian distribution of the end-to-end distance of the LRH1xC ensemble. A single narrow Gaussian distribution is an analogy to an ensemble composed of nearly 100% helical conformers. I used the predicted length (from Roy's model) for entirely helical segments and for fully extended chains as the lower and upper boundary. 40 Å is the length for an entirely helical LRH1xC. For fully extended chains, the length for a 26-residue peptide is around 80 Å. The mean length for a real LRH1xC chain should be between these two values. If we move the two labels along the chain, this number will change proportionally. The simulation follows the mathematical description of the Three-Energy Strategy and the Debye equation in Chapter II.

The features of the ASAXS curve change according to the shape of the distribution. For instance, if a Gaussian distribution is assumed, then the ASAXS curve is sensitive to both the mean and the variance (Fig.22). Generally speaking, the narrower the distribution is, the more oscillation patterns we can see in the resultant ASAXS curve, which is consistent with the Debye equation.

The distinct features, especially the position of the peaks and wells together with the height/depth of them, all contain useful information of the ensemble. But we must

realize that this is only from a single Gaussian simulation. It could be the case that when more and more distance distributions are added together, the ASAXS curve will become flattened and the distinct features will fade away, leaving us with a smooth and featureless curve. If this is true, then this method will not be very useful for studying unfolded ensembles since they often contain a broad range of conformers, which corresponds well to the case where distance distributions are combined to form a new broad distribution. With this in mind, I simulated the ASAXS signal using multiple Gaussian distributions.



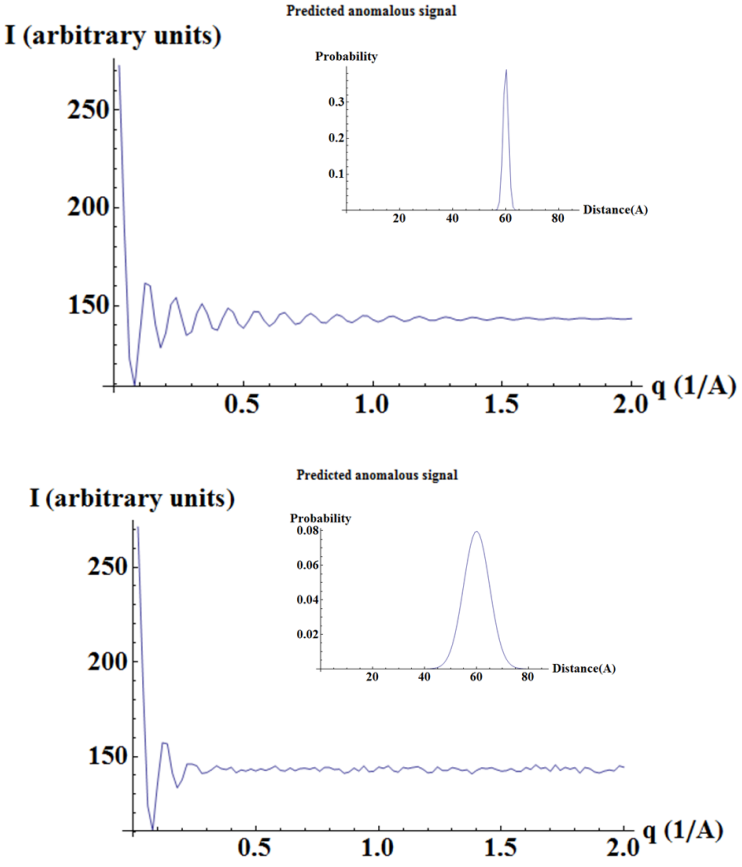


Figure 22: Simulated ASAXS curves from single Gaussian distributions of LRH1xC end-to-end distance. The four distributions are (first to last): μ = 40 Å, σ^2 = 1 Å^2; μ = 40 Å, σ^2 = 5 Å^2; μ = 60 Å, σ^2 = 1 Å^2; μ = 60 Å, σ^2 = 5 Å^2. Gaussian noise was added to the simulation.

4.2 Simulation of ASAXS signal using multiple Gaussian distance distribution

To mimic more realistic unfolded ensemble s, I devised the following steps:

(1) Generate two helix-coil ensembles that have average helicities of 10% and 70% respectively, since these are the minimum and maximum helicities of LRH1xC observed in the CD experiments performed by Jonathan. The mean distances for 0% and

100% helicity are 40 Å and 59 Å respectively, so those of the 10% and 70% distributions are 53.1 Å and 45.7 Å assuming that the mean distance changes linearly with helicity. This is an over-simplified assumption and it is to be improved when we establish the HCPP model in the future.

(2) In each case, for the 90% most populated conformers, compute their end-to-end distances based on appropriate polymer physics models.

(3) Compute the variance of end-to-end distance distribution for each conformer.

(4) Sum up these distributions to generate the end-to-end distance distribution of the ensemble.

(5) Use this distance distribution to compute the ASAXS signal with a 1/20 Gaussian noise level estimated from previous ASAXS data.

At present, we do not have a mature polymer physics model to carry out step (2) and (3). Alternatively, I used a random number generator to simulate the variances of the ensemble members, where the range is 0 to 10 for the high helicity ensemble and 10 to 20 for the low helicity ensemble, since the latter is more likely to assume a broad distribution. The ASAXS curve simulated from the 10% and 70% helicity ensembles are shown in Fig.23.

From the simulation results, we can see that the ASAXS curve is still sensitive to the shape of the distribution. We can also see that 0.05 $Å^{-1}$ to 0.6 $Å^{-1}$ is the q range where most important features occur, while the signal weakens significantly when q increases

above 0.6 $Å^{-1}$. This suggests to us the q range we should choose during future ASAXS experiments with the mercury label.

In order to know whether this degree of sensitivity is strong enough, we need to take into account the noise level in these experiments. Within this q range (0.05 $Å^{-1}$ to 0.6 $Å^{-1}$), the best SNR in the ASAXS curve we previous obtained was around 1/20. In the simulated curves, the difference intensity between two curves is mostly between 1/5 and 4 times the average intensity of the two, especially in the low q range (0.05-1) where the difference intensity can be as high as ten times the average intensity, which suggests that the signal is distinguishable from noise. The SNR of 1/20 was obtained from a sample concentration of 1 mg/mL, which suggests that even better SNR can be achieved by increasing the sample concentration to higher value.

However, we must keep in mind that although the ASAXS curve changes to the shape of distance distribution, the decomposition of the curve is still an ill-posed problem, meaning there could be more than one distance distribution solution to a given ASAXS curve. This is an innate nature of the SAXS data rather than a consequence of poor SNR, although a better SNR does increase the information content within an ASAXS curve, therefore putting more constraints on the decomposition result. In other words, it is impossible for us to determine the distance distribution of a given ensemble from a single ASAXS curve. To solve this problem, we need to mercury-label the peptide at different positions along the chain and measure many residue-to-residue scattering

curves. Mathematically speaking, this provides us with more equations to work with towards solving the inter-residue distance distribution of LRH1xC. At that time, we can train the HCPP model with the information obtained from our ASAXS experiments as explained in Chapter I.

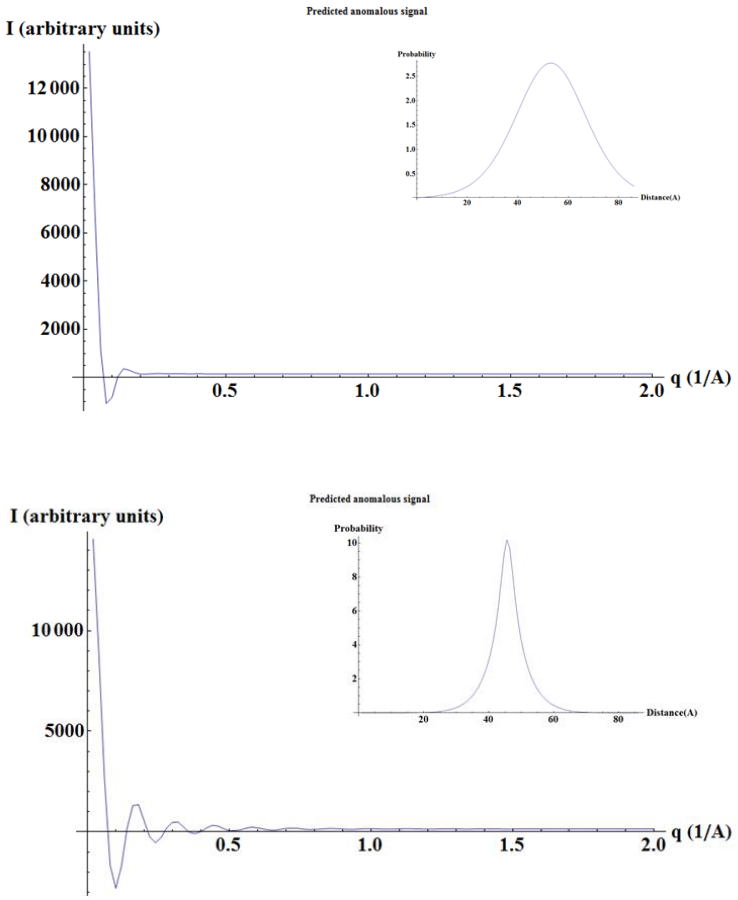


Figure 23: Simulated ASAXS curves from the combination of multiple Gaussian distributions of LRH1xC end-to-end distance. Upper: simulated ASAXS curve from the 10% helicity ensemble. Lower: simulated ASAXS curve from the 70% helicity ensemble.

Conclusions

I demonstrated that the residue-to-residue distance distribution of the unfolded protein ensemble can be measured via a small-angle X-ray scattering technique. The two residues of interest are labeled with selenium or mercury and the distance distribution is extracted from the anomalous scattering signal using the Three-Energy Strategy. By simulation, we can see the anomalous scattering signal contains useful information of the ensemble. End-to-end distance distribution of LRH1xC was obtained using the selenium label. The result shows that a broad distribution with the maximum population at 8 Å, which is rather unexpected. The signal-to-noise ratio is low in mercury-labeled LRH1xC and no further data processing was carried out. Better signal-to-noise ratio can be achieved by increasing the sample concentration of mercury-labeled LRH1xC in the future.

References

1. Bartels T, Choi JC and Selkoe DJ. α-synuclein occurs physiologically as a helically folded tetramer that resists aggregation. Nature **477**, 107-111 (2011)

2. Schmidler SC, Lucas JE. and Oas TG. Statistical estimation of statistical mechanical models: helix-coil theory and peptide helicity prediction. Journal of Computational Biology **14**, 1287-1310 (2007)

3. Zhou H. Polymer models of protein stability, folding, and interactions. Biochemistry **43**, 2141-2154 (2004)

4. Rubinstein M and Colby R. Polymer Physics. Oxford University Press (2003)

5. Beamer L and Pabo C. Refined 1.8 Å crystal structure of the λ repressor-operator complex. Journal of Molecular Biology **227**, 177-196 (1992)

6. Huang G and Oas TG. Structure and stability of monomeric λ repressor: NMR evidence for two-state folding. Biochemistry **34**, 3884-3892 (1995)

7. Muller S, Senn H, Gsell B, Vetter W, Baron C and Bock A. The formation of diselenide bridges in proteins by incorporation of selenocysteine residues: biosynthesis and characterization of $(Se)_2$-thioredoxin. Biochemistry **343**, 3404-3412 (1994)

8. Trumpler S, Lohmann W, Meermann B, Buscher W, Sperling M and Karst U. Interaction of thimerosal with proteins-ethylmercury adduct formation of human serum albumin and β-lactoglobulin A. Metallomics **1**, 87-91 (2009)

9. Putnam CD, Hammel M, Hura GL. and Tainer JA. X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Quarterly Reviews of Biophysics **40**, 191-285 (2007)

10. Hendrickson W. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science **254**, 51-58 (1991)

11. Cantor C and Schimmel P. Biophysical chemistry. W. H. Freeman & Co., San Francisco (1980)

12. Jalilehvand F. Lecture materials.

13. Goerigk G, Schweins R, Huber K and Ballauff M. The distribution of Sr^{2+} counterions around polyacrylate chains analyzed by anomalous small-angle X-ray scattering. Europhysics Letters **66**, 331-337 (2004)

14. Teixeira J. Small-angle scattering by fractal systems. Journal of Applied Crystallography **21**, 781-785 (1988)

15. Johansen D, Trewhella J, Goldenberg DP. Fractal dimension of an intrinsically disordered protein: small-angle X-ray scattering and computational study of the bacteriophage λ N protein. Protein Science **20**, 1955-1970 (2011)

16. Kohn JE, Millett IS, Jacob J, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, Ruczinski I, Doniach S, Plaxco KW. Random-coil behavior and the dimensions of chemically unfolded proteins. Proceedings of the National Academy of Sciences **101**, 12491-12496 (2004)